# A Comparative Evaluation of Multiple Chat Stream Interfaces for Information-intensive Environments

**Yiran Wang, Andy Echenique, Martin Shelton, Gloria Mark**
Department of Informatics
University of California, Irvine
{yiranw2, echeniqa, mshelton, gmark@uci.edu}

**ABSTRACT**
For information workers who monitor numerous constantly updating data streams, conserving cognitive resources is crucial. This study evaluated how an interface affects information workers' ability to grasp critical information from multiple text-based chat streams under time pressure. We designed and built a working prototype that displays ten chat streams simultaneously in standard chat windows (ST) and ticker tapes (TT). We conducted a lab experiment to evaluate differences in how these two interfaces support signal and context detection. We found that with ST, participants detected significantly more target words (SAT words) with rarer frequency and significantly more context information (disaster facts) than with TT. Our results show that while TT is potentially better for overview scanning of multiple streams, ST is likely to be better for multi-tasking. Our study informs the design of future multi-chat systems so that large amounts of information can be easier to detect and process.

**Author Keywords**
Monitoring multiple text-based streams; interface evaluation; information workers; empirical study.

**ACM Classification Keywords**
H.5.2 [Information Interfaces and Presentation]: User Interfaces – Screen design; Evaluation/methodology.

## INTRODUCTION

Information workers who deal with time-sensitive and context-dependent information, such as intelligence or business analysts, must constantly monitor a large number of real-time data streams in order to discover events or make trading decisions. For this type of work, information overload can hinder critical information detection and situational comprehension. As [3] summarizes: "attention pressures are intensifying as operators must regularly multitask, accommodate frequent interruptions, and attend to numerous information sources". Because humans can attend to only a small amount of information on a visual display at one time, the design of how multiple data streams are displayed on interfaces could affect how people allocate their attention.

In addition to attentional constraints, there are also the limitations of human memory. Tulving and Thomson [7] suggest that the process of encoding and storing information into memory begins as soon as the stimulus is perceived and attended to. However, when attention is divided amongst concurrent tasks, encoding performance can deteriorate [2]. In multiple stream chat environments, participants are seldom allowed time for undivided attention, straining attentional resources.

A number of interfaces have been designed to display text-based chats, such as thread-based chats, ticker tape chats, spatially arranged chats with avatars [5] and motion-based chats [1]. As Roddy and Epelman-Wang [5] pointed out, the long unsolved problem with text-based chat systems is that they are not well-designed for monitoring multiple, concurrent conversations. While a scalability concern is based on the increasing number of users in one chat window, it is also important to look at the issue of increasing numbers of chat windows. For example, close observations of experienced operators in a command and control environment revealed they can have on average six to eight, and up to 14, chat windows open on one screen while constantly shifting among active and inactive windows [3]. Thus, information workers like these operators face numerous challenges to attend to many chat windows in parallel, especially in light of distractions, memory overload and prioritization [3]. The purpose of this study is to evaluate how different interface designs can support information workers' ability to grasp relevant and timely information from multiple active chat streams.

## INTERFACE DESIGNS FOR MULTIPLE CHAT STREAMS

We designed and built working prototypes of two interfaces to display multiple chat streams: a standard chat window (**ST**) and a ticker tape (**TT**) (see Figure 1). We chose TT as an alternative to compare with ST because each represents different sets of design features:


*CHI'13,* April 27 – May 2, 2013, Paris, France.

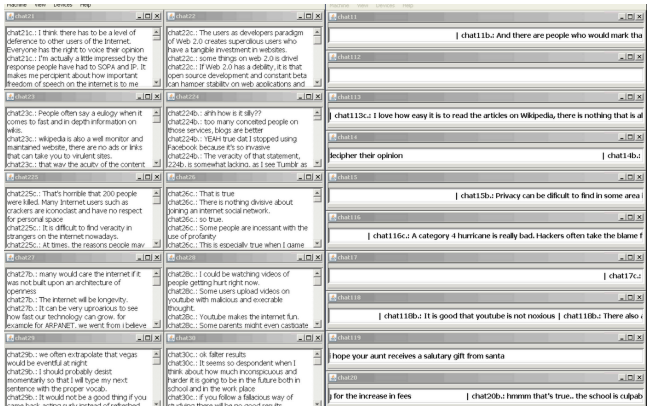

**Figure 1. Interface prototype of ST (left) and TT (right)**

1) **Screen real estate**: In ST, multiple square windows are arranged in a 2x5 matrix whereas in TT, multiple chat windows are stacked vertically. Because of its single-lined feature, the TT interface takes up minimal screen space [4].

2) **Text updating orientation**: With ST, new messages abruptly appear as a whole sentence from the top of each chat window and push the older text downwards or out of the window immediately. With TT, incoming text streams flow continuously from right to left which gives users temporally relevant information at a focal point [4].

3) **Text history**: In ST, if no new text appears, a history of the dialogue remains in the window. Conversely, all inactive windows are empty in TT.

In sum, text is distributed across separate areas of the whole screen in ST while with TT, content moves across the screen. We believe that users could potentially scan more information from TT with a lower cognitive load by focusing on the center of the screen as information streams by, as opposed to scanning the whole screen with ST. However, because the text is constantly moving off the screen, we think ST might be better for grasping context because it (temporarily) maintains more chat history. We wanted to investigate which motion type is better suited for information detection.

Prior studies have focused on the usage of one TT window, rather than many. In a previous study [4], TT was used in an awareness widget design to provide users with awareness of events in distributed work environments. According to [4], the tickertape interface was accepted well among users based on their adoption. TT was also used as a chat interface that supports multiple users, demonstrating the potential to address the problem of scalability of number of users per window, focus, and responsiveness [5]. However, none of these studies using the TT design experimentally tested the efficiency of the interface. Our study differs from previous studies in that 1) we empirically evaluate signal and context detection supported by each interface as opposed to self-reported acceptance or preference; 2) we compare multiple chat windows as opposed to a single window; and 3) we test the interfaces with numerous continuous chat streams, which is a similar environment to what analysts experience.

| | Word Frequency | | | | | | Sum |
|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | |
| Corpus A | 85 | 1 | 30 | 1 | 13 | 17 | 147 |
| Corpus B | 87 | 2 | 25 | 7 | 17 | 11 | 149 |
| Collected(All) | 172 | 3 | 55 | 8 | 30 | 28 | 296 |
| **Designed(All)** | **180** | **0** | **60** | **0** | **36** | **30** | **306** |

**Table 1. Distribution of signal frequency**

## METHOD

### Experimental Stimuli

We first collected a chat corpus for our experimental task from 60 undergraduate student volunteers to simulate complex, information-rich environments containing specific words and phrases of interest. Students were grouped into 20 chat rooms (three per room). They were asked to chat about an assigned topic and use 12 given SAT words of medium rarity (*e.g.* "*demeanor*", "*enigmatic*") [6]. The data collection resulted in 20 chat room conversations.

The data were divided into two corpuses (A and B), each with ten chat rooms and 153 unique SAT words. Each SAT word's frequency was varied in order to have different levels of detection difficulty, similar to what analysts would experience: e.g. rare signals such as keywords to indicate terrorist plots amidst more frequent keywords. To simulate such variability in signals, we designed for words to appear once, three times, five times, or six times. However, errors during collection resulted in slightly uneven distributions; word frequencies of two and four in the corpuses collected were due to errors. These inconsistencies were minor and we believe that the corpuses are comparable (see Table 1).

In order to assess comprehension of content, we added additional information into the chat rooms. Seven facts relating to a natural disaster were inserted into each corpus. These facts were comparable in each corpus (*e.g.* "The hurricane already hit Houston" was embedded in corpus A; "The earthquake already hit San Francisco" was embedded in corpus B).

#### *Chat Display*

We displayed each corpus on the ST and TT interfaces. Both interfaces displayed ten chat streams in ten chat windows simultaneously. Both ST and TT text were irretrievable after scrolling off the screen since we did not design a "scroll bar" in the experiment. The chat playback was then video recorded and used as experimental stimuli. In each session, we displayed the chat videos on the left half of a 13-inch laptop screen (1280x800 resolution), with ten 2.85x1.42 inch ST windows or 5.71x0.71 inch TT windows. An empty notepad was displayed on the right half of the screen for note taking. The length for each video ranged from 10 minutes 46 seconds to 11 minutes. Each

word moved across the width of the TT window in 4.7 seconds, and ST had seven lines of text per window. We also recorded a five-minute video containing no SAT words for training before each task, comparable to the actual task conditions.

### Participants

We had 18 undergraduate student participants majoring in computer science fields, from a university on the U.S. west coast. Due to two experimental errors by researchers and task failure by one participant, data from 3 participants are excluded in our analyses. In total, we collected valid data from 15 participants: 4 females and 11 males.

### Procedure

We used a within-subjects design: each participant was instructed to perform a task on both TT and ST. The presentation order of the interfaces was counter-balanced. Before each task, participants were given a five-minute training task to acclimate them with the corresponding interface.

In each task, participants were instructed to watch the chat video and simultaneously type into the empty notepad all uncommon words that they believed to be SAT words (with repeats if any). They were also asked to remember (without writing down) facts about significant natural disasters in the chat streams. After each task, participants were asked to copy the SAT words (with the number of times they appeared) from the notepad into an online form together with the disaster facts they remembered. Finally, a post-experimental survey was given to collect interface preference and strategies used during the task. The whole experimental session averaged 1 hour 10 minutes, ranging from 59 minutes to 1 hour 24 minutes.

## RESULTS

We evaluated task performance in two ways: signal detection, based on the SAT target words recorded and context, through the natural disaster facts recorded. We evaluated the usability of each interface based on the post-experimental survey.

### SAT Words and Context Detection

We compared the word detection in each interface. The total value for word detection was weighted according to the SAT word counts participants reported. Though the total number of target words noted in ST (mean=60.20, SD=26.96) was higher than those in TT (mean=57.00, SD=23.78), a paired t-test showed this was not a significant difference ($t=1.03$, $p=0.32$).

The number of unique target words was also compared between interfaces. The number of unique words was calculated by counting each correct SAT word response in the task only once, regardless of the number of same words reported. A paired t-test showed a trend of ST having a larger average unique word count (41.73, SD=12.99) than TT (37.67, SD=14.47), ($t=1.98$, $p<.07$).

Word frequencies were compared between interfaces. The differences would demonstrate if certain word rarities are easier to detect on an interface. Words that appeared once (singles) were recorded significantly more for ST (14.08, SD=5.60) than TT (12.06, SD=4.93), ($t=2.93$, $p<.02$). No significant difference was found for other word frequencies (word frequency of 3: $t=0.71$, $p=0.49$; word frequency of 5: $t=0.56$, $p=0.58$; and word frequency of 6: $t=1.15$, $p=0.27$). Subjects thus had better performance with ST for detecting words that appeared once.

The detection of contextual information (i.e. facts related to natural disasters) was also compared between interfaces. These were calculated by counting how many of the seven facts each participant reported correctly. ST (mean=2.8, SD=1.61) was again found to have a significantly higher detection rate than TT (mean=1.67, SD=1.18), ($t=2.20$, $p<.04$).

These results show that for detection of single frequency words and context facts, ST had a statistically significant higher detection rate than TT. Because target word detection involved two stages: scanning chat streams and typing down target words (which simulated multi-tasking), we next investigated which of the two stages ST was helping. In order to give further insight to this question, we analyzed the post-experimental survey results.

### Post-experimental Survey Results

Surveys were completed at the end of the study. Participants were asked to rate their preference between ST and TT. Out of the 11 collected responses, seven preferred ST as their overall preferred interface. While eight out of nine participants reported ST to be easier to type while reading, seven out of 11 reported TT easier to read. The preference ratings and the task performances together indicate that participants may have performed better with ST because it better assisted multi-tasking (typing while reading).

Further supporting the notion that ST is better for multitasking (supporting typing while reading) is seen in the strategies participants revealed in the open-ended questions. The most common strategy mentioned by five participants with ST was how they kept track of newly updated chat windows. While typing, they could still see what chats became updated by looking for movement in the chat windows. Once they were done writing, they could then read the newly updated chat windows and scan for target words. As participant A13 notes, *"[I]…mostly wait for a chat to move... would scan top to bottom sometimes, but then switched to just wait for motion..."* These results support the idea that the standard chat is better than tickertape concerning how well the interface is paired with activities of scanning and typing – the multi-tasking that intelligence analysts need to perform.

*Pros and Cons for Standard Chat and Ticker Tape*

The finals comments indicate no universal preference for one interface. They do, however, reveal a set of pros and cons for ST and TT interfaces. Some repeated comments for ST are: *easier to read conversation because [the words] stayed still, not disorienting, and easier to read history*. However, they were also concerned that they missed parts of the conversation. For example, as participant B16 mentioned: "if things pop up at the same time, I'll just pick one and ignore the rest". This indicates that while ST is good for checking history when the volume of incoming text is low, it is not good for overall scanning if many chat windows are active. Additionally, several participants mentioned the problem that new lines of chat bump the old line off the window immediately if there are multiple lines.

Participants reported that TT *was easier to scan, easier to look at the entire screen, and easier to monitor different windows simultaneously*. However, as four participants mentioned, the most dominant disadvantage is that the constant motion of text caused physical discomfort such as disorientation, dizziness or headaches.

## DISCUSSIONS AND CONCLUSIONS

Our experiment showed that ST supported higher target detection for words appearing once, indicating it is favorable for detecting uncommon signals. We therefore suggest ST for work environments where analysts must detect unusual signals or abnormal events while multitasking. Our study did not test whether ST better leverages spatial memory, i.e. the memory of chat streams based on their screen position. We think ST might have an advantage over TT in utilizing spatial memory since it could be hard to differentiate positions of certain text in the TT horizontal stack of single chat lines.

We stress that target word detection performance measured subjects' constant task switching between detection and typing (noting down the signals). Because ST maintains chat history for inactive windows while the sudden motions from incoming chat make the active windows easy to notice, we believe it is very likely that ST supports better multitasking. Participants could type or visit chat history and instantly assess the newly updated window for potential new signals, since "waiting for the motion of updates" was the participant's most common strategy. Whereas with TT, in order for participants to switch from notepad to the chat streams, they had to constantly divert attention to finding where they left off, which was more of a cognitive load.

Our experiment also showed that ST assisted significantly higher detection based on context (detection of natural disaster facts). We suggest two reasons: 1) the chat history in inactive windows made it easy for participants to grasp contextual information; and 2) participants tended to read with ST as opposed to skimming with TT, as they reported in their strategies. The opportunity to read (and process) text facilitates context understanding.

Although participants' task performances were worse with TT, our qualitative results suggested that TT might be better for scanning the whole screen. Out of 11 participants, seven rated TT to be easier to read; five commented that they would scan the whole screen up and down; five reported that they could scan multiple tickers at the same time. This suggests that TT supports overview awareness as multiple chats steadily pass by. The reason why the "easier to scan" feature failed to result in better performance may be due to the dizziness and disorientation caused by the constant text motion. We thus suggest that information workers who need to monitor data streams for long hours to avoid using an interface with steady and continuous texts flow.

In this study, we conducted a lab experiment to evaluate two text-based chat interfaces (ST and TT) on how they assist information workers to detect key information from multiple chat streams in a time sensitive context. Other approaches could utilize the psychological notion of word priming or an interface feature such as highlighting. We feel our results could inform the design of large-scale text information visualizations for information workers who monitor real-time text streams.

## ACKNOWLEDGMENTS

We thank Steve Voida and Judy Olson for their invaluable help. This project was supported by NSF grant #0910640 and Northrup Grumman.

## REFERENCES

1. Albrecht-Buehler, C., Watson, B. and Shamma, D. A. Visualizing Live Text Streams Using Motion and Temporal Pooling. *Computer Graphics and Applications*, IEEE 25, 3 (2005), 52-59.
2. Bargh, J. A. The four horsemen of automaticity: Awareness, efficiency, attention, and control in social cognition. In R.S. Wyer and T.K. Srull (eds). *Handbook of Social Cognition*, Hillsdale, NJ, 1994, 1-40.
3. Boiney, L. G., Goodman, B., Gaimari, R., Zarrella, J., Berube, C. and Hitzeman, J. Taming Multiple Chat Room Collaboration: Real-time Visual Cues to Social Networks and Emerging Threads. In *Proc. ISCRAM* (2008), 660-668.
4. Parsowith, S., Fitzpatrick, G., Kaplan, S., Segall, B. and Boot, J. Tickertape: Notification and Communication in a Single Line. In *Proc. CHI*, (1998), 139-144.
5. Roddy, B. J. and Epelman-Wang, H. Interface Issues in Text Based Chat Rooms. *SIGCHI Bull.* 30, 2 (1998), 119–123.
6. SubtlexUS: American Word Frequencies. http://subtlexus.lexique.org/
7. Tulving, E. and Thomson, D. M. Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, 80 (1973), 352-373.